Information Extraction
-
Architecture and Task Definition

PD Dr. Günter Neumann
DFKI and Saarland University

Next Lecture at 10th Sept

- Geb. C7.2 Seminarraum

Information Extraction (IE)

The goal of IE research is to build systems that find and link *relevant* information from NL text ignoring irrelevant information.

Core Functionality

Input

Templates coding relevant information, e.g. company, product, medical information
set of real world texts

Output

set of instantiated templates filled with relevant text fragments (normalized to a canonical form)

Example: Job Advertisment

Input

Posting from Newsgroup

```
Telecommunications. SOLARIS Systems
Administrator. 38-44K. Immediate need

Leading telecommunications firm in need
of an energetic individual to fill the
following position in the Atlanta
office:

  SOLARIS SYSTEMS ADMINISTRATOR
  Salary: 38-44K with full benefits
  Location: Atlanta Georgia, no
        relocation assistance provided
```

Output

Filled Template

```
computer_science_job
title: SOLARIS Systems Administrator
salary: 38-44K
state: Georgia
city: Atlanta
platform: SOLARIS
area: telecommunications
```

Example: Terrorists actions

„Salvadoran President-elect Afredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of crime.“

Incident:	KILLING
Perpetrator:	„terrorist“
Confidence:	__
Human Target:	„Roberto Garcia Alvarado“

Incident:	KILLING
Perpetrator:	FMLN
Confidence:	Accused by Authorities
Human Target:	__

Incident:	KILLING
Perpetrator:	FMLN
Confidence:	Accused by Authorities
Human Target:	„Roberto Garcia Alvarado

Example: Company‘s turnover

Lübeck (dpa) – Die Lübecker Possehl-Gruppe, ein im Produktions-, Handel- und Dienstleistungsbereich tätiger Mischkonzern, hat 1994 den Umsatz kräftig um 17 Prozent auf rund 2,8 Milliarden DM gesteigert. In das neue Geschäftsjahr sei man ebenfalls „mit Schwung“ gestartet. Im 1. Halbjahr 1995 hätten sich die Umsätze des Konzerns im Vergleich zur Vorjahresperiode um fast 23 Prozent auf rund 1,3 Milliarden erhöht.

Type:	turnover
C-name:	Possehl1
Year:	1994
Amount:	2.8e+9DM
Tendency:	+
Diff:	+17%

Type:	turnover
C-name:	Possehl1
Year:	1995/1
Amount:	1.3e+9DM
Tendency:	+
Diff:	+23%

Text-based Information Extraction (IE)

Template:

ManagementSuccession
PersonIn: _____
PersonOut: _____
Position: _____
Organisation: _____
TimeIn: _____
TimeOut: _____

documents

Dr. Hermann Wirth, bisheriger **Leiter** der **Musikhochschule München**, verabschiedete sich heute aus dem Amt. Der 65jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde **Sabine Klinger** benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.

Text classification

Linguistic processing

Template processing

Linguistic processing
- tokenization
- morphology
- chunks
- Clause topology
- Gram. functions
- Reference-resolution

Template processing
- Domain lexicon
- Named Entities
- Relations
- Events

ManagementSuccession
PersonIn: *Klinger*
PersonOut: *Wirth*
Position: *Leiter*
Organisation: *Musikhochschule München*
TimeIn: _____
TimeOut: *3.4.2002*

Example: LIEP (S. Huffman, 1995)

<PNG> Sue Smith </PNG>, 39, of Menlo Park, was appointed <TNG> president </TNG> of <CNG> Foo Inc. </CNG>

n_was_named_t_by_c:

noun-group(PNG, head(isa(person-name))),

noun-group(TNG, head(isa(title))),

noun-group(CNG,head(isa(company-name))),

verb-group(VG, type(passive), head(named or elected or appointed)),

prep(PREP, head(of or at or by)),

subject(PNG,VG), object(VG,TNG), post_nominal_prep(TNG,PREP), prep_obj(PREP,CNG)

⇒ management_appointment(M,person(PNG), title(TNG), company(CNG))

Major IE tasks

- Named Entity task (NE)
- Template Element task (TE)
- Template Relation task (TR)
- Scenario Template task (ST)
- Co-reference task (CO)

Named Entity Task (NE)

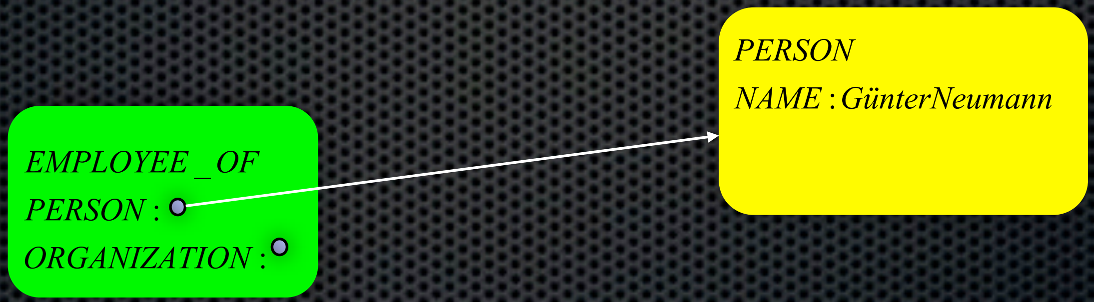
Mark into the text each string that represents a person, organization, or location name, or a date or time, or a currency or percentage figure (this classification of NEs reflects the standard types of NE applied in IE).

$PERSON$

$NAME: GünterNeumann$

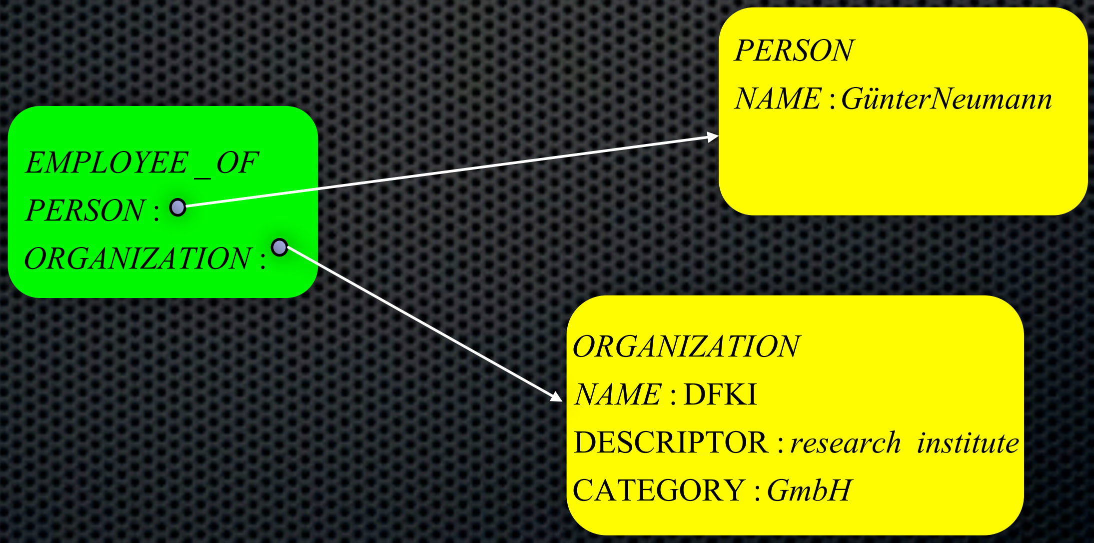
Template Element Task (TE)

Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (also known as slot filler task; it is basically a binary relation of form „attribute of x has value y“)



Template Relation task (TR)

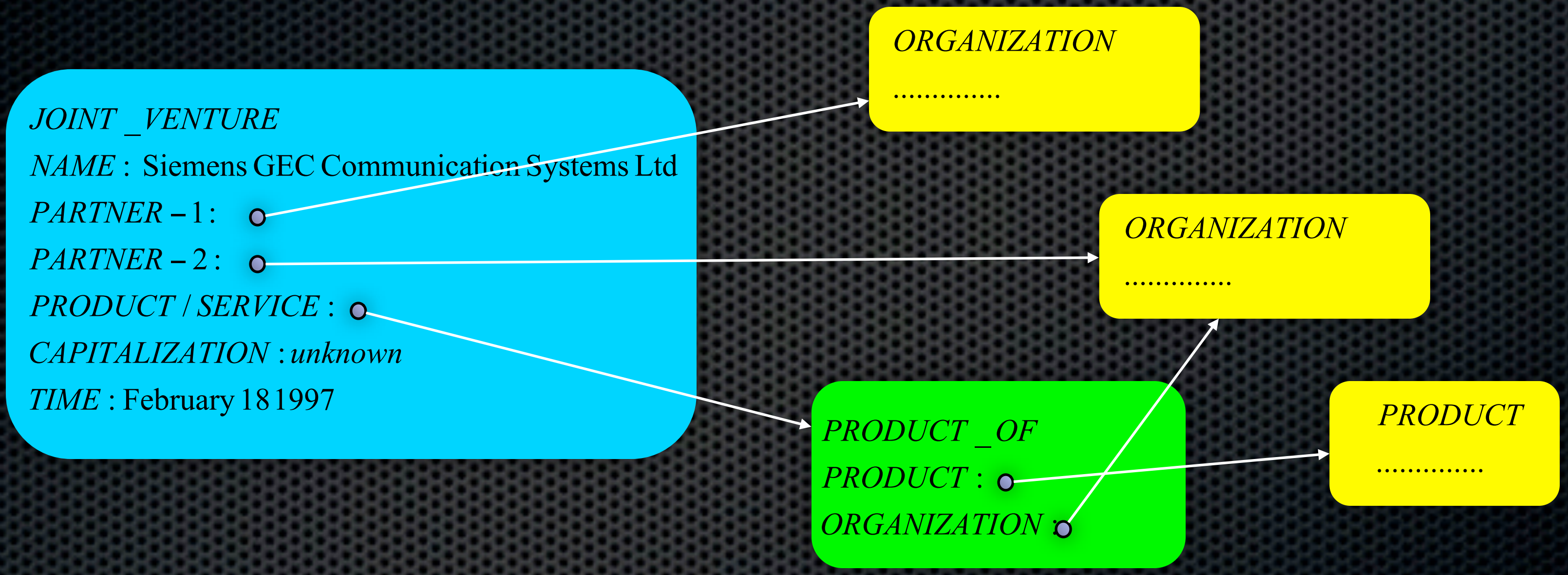
Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain-independent relationships between entities identified by TE)



Scenario Template task (ST)

Extract pre-specified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)

ST example


JOINT _VENTURE
NAME : Siemens GEC Communication Systems Ltd
PARTNER – 1 :
PARTNER – 2 :
PRODUCT / SERVICE :
CAPITALIZATION : unknown
TIME : February 18 1997
ORGANIZATION
....
ORGANIZATION
....
PRODUCT _OF
PRODUCT :
ORGANIZATION :
PRODUCT
....

Coreference task (CO)

Capture information on co-referring expressions, i.e. all mentions of a given entity, including those marked in NE and TE (Nouns, Noun phrases, Pronouns)

He

The Artist Formerly Known As Prince

Prince

An Example

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

- NE: *red rocket*, *Tuesday*, *Dr. Big Head, Dr. Head,* and *We Build Rockets Inc.*
- CO: *it* refers to the rocket; *Dr. Head* and *Dr. Big Head* are the same
- TE: the rocket is *shiny red* and Head's *brainchild*
- TR: Dr. Head *works for* We Build Rockets Inc.
- ST: a *rocket launching event* occurred with the various participants.

Scoring templates

- Templates are compared on a slot-by-slot basis
  - Correct: response = key
  - Partial: response ≈ key
  - Incorrect: response ≠ key
  - Spurious: key is blank
    - overgen=spurious/actual
  - Missing: response is blank

Evaluation Metrics

- Precision and recall:
  - Precision: correct answers/answers produced
  - Recall: correct answers/total possible answers
- F-measure
  - Where β is a parameter representing relative importance of P & R:

$$F = \frac{(\beta^2 + 1)PR}{(\beta^2 P + R)}$$

  - E.g., β=1, then P&R equal weight, β=0, then only P
- Current State-of-Art: F=.60 barrier

Maximum Results Reported in MUC-7 (Message Understanding Conference, 2001)

Measure\Task	NE	CO	TE	TR	ST
Recall	92	56	86	67	42
Precision	95	69	87	86	65

Human on NE task	F	R	P
Annotator 1	98.6	98	98
Annotator 2	96.9	96	98

Human on ST task: ~ 80 % F

This is mainly for hand-written systems

Jerry Hobbs: Why the 60% Barrier

1. Merging problems accounted for 60% of our errors.

2. Entity recognition performance is 90%; event recognition requires recognizing ~4 entities; $.9^4 = .6$

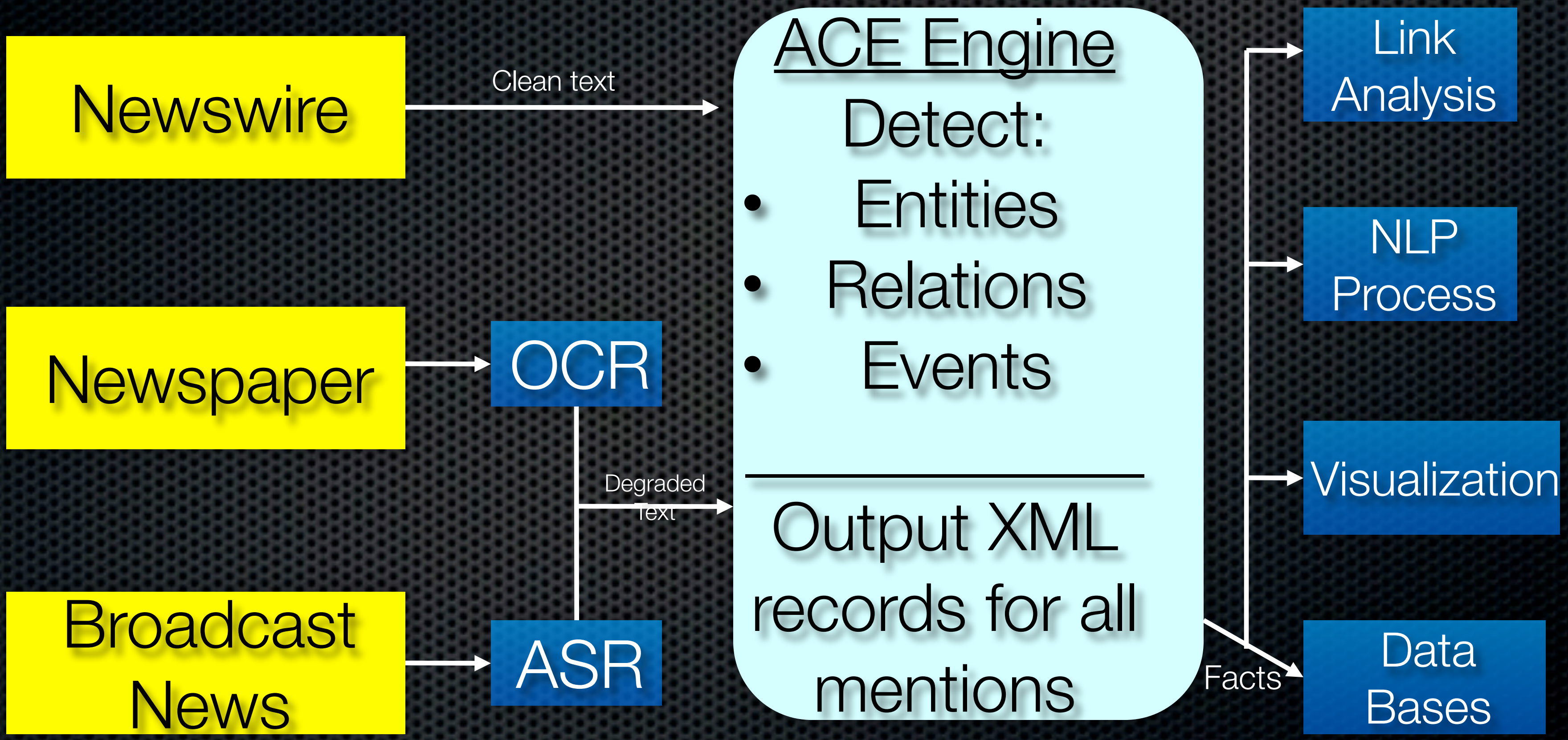
3. The distribution of problems has a very long tail.

4. 60% is what the text wears on its sleeve; the rest is implicit and requires inference and world knowledge.

Automatic Content Extraction - ACE (cf. Appelt, 2003)

- Develop core information extraction technology by focusing on extracting specific semantic entities and relations over a very wide range of texts.
- Corpora: Newswire and broadcast transcripts, but broad range of topics and genres.
  - Third person reports
  - Interviews
  - Editorials
  - Topics: foreign relations, significant events, human interest, sports, weather
- Discourage highly domain- and genre-dependent solutions

The Technical Approach



Objectives and Performance Goals of ACE

- Objectives
  - Extract Info from Texts of Varying Quality
  - Detect Unique entities, relations, events
    - Find all mentions within documents
    - Collect all mentions by object
  - Track entities within & across documents
  - Output XML for follow-on processes
- Performance Goals
  - Extract 95% of the value in document

Example for entities & their mentions

[COLOGNE, [Germany]] (AP) _ [A [Chilean] exile] has filed a complaint against [former [Chilean] dictator Gen. Augusto Pinochet] accusing [him] of responsibility for [her] arrest and torture in [Chile] in 1973, [prosecutors] said Tuesday.
[The woman, [[a Chilean] who has since gained [German] citizenship]], accused [Pinochet] of depriving [her] of personal liberty and causing bodily harm during [her] arrest and torture.

Person
Organization
Geopolitical Entity

Core Mission: Information Gathering

- Information content is main interest of human language text
  - Semantics drives information gathering
  - Syntax is the vehicle for organizing the information
- ACE systems provide NL understanding
  - Detect each entity, relation, and event of specific type
  - Recognize all mentions of entities, relations & events
  - Resolve all mentions to the proper entity, relation, or event
- Convert information in human language into structured data
  - Extract semantics of communication
  - Output in ACE program format
- Structured data supports real world modeling & analysis

Components of a Semantic Model

- Entities - Individuals in the world *that are mentioned in a text*
  - Simple entities: singular objects
  - Collective entities: sets of objects of the same type *where the set is explicitly mentioned in the text*
- Attributes - Timeless unary properties of entities (e.g. Name)
- Temporal points and intervals
- Relations - Properties that hold of two entities over a time interval
- Events - A particular kind of relation among entities implying a change in relation state at the end of the time interval.

Semantic Analysis: Relating Language to the Model

- Linguistic Mention
  - A particular linguistic phrase
  - Denotes a particular entity, relation, or event
    - A noun phrase, name, or possessive pronoun
    - A verb, nominalization, compound nominal, or other linguistic construct relating other linguistic mentions
- Linguistic Entity
  - Equivalence class of mentions with same meaning
    - Co-referring noun phrases
    - Relations and events derived from different mentions, but conveying the same meaning

Choosing an Ontology for IE Semantics

- Ordinary native speakers should be able to annotate text with minimal training.
- People should have well-developed intuitions about type classification
  - Is a “museum” an organization or facility?
- People should have well-developed intuitions about entity coreference
  - “Peace in the Middle East”
- Entities should be extensional, not abstract, generic, counterfactual, or fictional

Relations

- Relations hold between two entities over a time interval.
- Relations may be “timeless” or temporal interval is not specified
- Relations have inertia, i. e. they don’t change unless a relevant event happens.

Explicit and Implicit Relations

- Many relations are true in the world. Reasonable knowledge bases used by extraction systems will include many of these relations. Semantic analysis requires focusing on certain ones that are directly motivated by the text.
- Example:
  - Baltimore is in Maryland is in United States.
  - “Baltimore, MD”
  - Text mentions Baltimore and United States. Is there a relation between Baltimore and United States?

Explicit Relations

- Explicit relations are expressed by certain surface linguistic forms
  - Copular predication - Clinton was the president.
  - Prepositional Phrase - The CEO of Microsoft…
  - Prenominal modification - The American envoy…
  - Possessive - Microsoft's chief scientist…
  - SVO relations - Clinton arrived in Tel Aviv…
  - Nominalizations - Anan's visit to Baghdad…
  - Apposition - [Tony Blair, [Britain's prime minister]…]

Textual Analysis Conference (since 2009) - Knowledge Base Population Track

Motivation

- **IE & QA technologies have been studied in isolation**
  - **Not focused on discovery of information for inclusion in an existing knowledge base**
  - **No consideration of novelty, contradiction**

- **Issues when filling in a KB**
  - **Accurate extraction of facts**
  - **Global resolution of entities**
  - **Maintaining provenance of asserted facts**
  - **Avoiding contradiction / detection of novel information**
  - **Temporal qualification of assertions**
  - **Leveraging existing KB to assist with extraction**
  - **Scalability**

16 November 2009

Textual Analysis Conference (since 2009) - Knowledge Base Population Track

TAC

Comparison to ACE & TREC-QA

- **Corpus vs. document focus**
  - ACE: component tasks (NER, relation extraction) for a set of isolated documents
  - KBP: learn facts from a corpus. Repetition not very important. Asserting wrong information is bad.
- **Context**
  - In KBP, there is a reference knowledge base, so avoiding redundancy and detecting contradiction are important
  - In KBP slots are fixed and targets change. In TREC QA, the targets dictated which questions were asked.
- **Knowing when you don't know**
  - TREC QA had a small percentage of NIL questions (4-10%)

16 November 2009

Textual Analysis Conference (since 2009) - Knowledge Base Population Track

TAC

KBP Snapshot

- **Track structure**
  - NIST – overall organization, infrastructure, evaluation
  - LDC – develop and distribute data resources, target selection, human assessments
- **Datasets**
  - LDC produced 1.3M English newswire collection
  - Reference KB populated with semi-structured facts obtained from English Wikipedia (Oct '08 dump)
    - 200k people, 200k GPEs, 60k orgs, 300+k misc/non-entities
- **Two tasks**
  - Entity Linking - Grounding entity mentions in documents to KB entries
  - Slot Filling - Learning attributes about target entities

16 November 2009

Textual Analysis Conference (since 2009) - Knowledge Base Population Track

Textual Analysis Conference (since 2009)
- Knowledge Base Population Track

TAC

Most Frequent KB Classes

95142 settlement
72992 album
34659 film
32464 musical artist
23138 actor
21195 single
16765 company
15644 book
14567 football biography
14121 person
12646 radio station
12514 nrhp
12324 vg
11813 planet
10818 uk place
10113 television

8353 ort in deutschland
8061 university
7675 airport
7492 military person
7270 road
7185 indian jurisdiction
7123 cityit
6143 australian place
6131 mountain
5957 military conflict
5952 military unit
5937 city
5630 software
5501 mlb retired
5397 writer
5349 scientist

5222 lake
4913 television episode
4636 school
4426 commune de france
4265 aircraft
4229 ice hockey player
3918 german location
3234 nflactive
3168 disease
3070 politician
3036 u.s. county
2956 station
2950 automobile
2933 officeholder
2833 broadcast
2728 swiss town

PER ORG GPE OTHER

16 November 2009

Textual Analysis Conference (since 2009) - Knowledge Base Population Track

Entity Linking Task

John Williams

Richard Kaufman goes a long way back with John Williams. Trained as a classical violinist, Californian Kaufman started doing session work in the Hollywood studios in the 1970s. One of his movies was Jaws, with Williams conducting his score in recording sessions in 1975...

John Williams	author	1922-1994
J. Lloyd Williams	botanist	1854-1945
John Williams	politician	1955-
John J. Williams	US Senator	1904-1988
John Williams	Archbishop	1582-1650
John Williams	composer	1932-
Jonathan Williams	poet	1929-

Michael Phelps

Debbie Phelps, the mother of swimming star Michael Phelps, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

Michael Phelps is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. Phelps has ...

Michael Phelps	swimmer	1985-
Michael Phelps	biophysicist	1939-

Identify matching entry, or determine that entity is missing from KB

16 November 2009

Textual Analysis Conference (since 2009)
- Knowledge Base Population Track

TAC

Slot Filling Task

Target: EPA
(plus 1 document)

Environmental Protection Agency

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

Agency overview	
Employees	17,964 (2005)
Annual budget	$7.3 billion (2007)
Agency executive	Lisa P. Jackson, Administrator

Generic Entity Classes

Person, Organization, GPE

Missing information to mine from text:

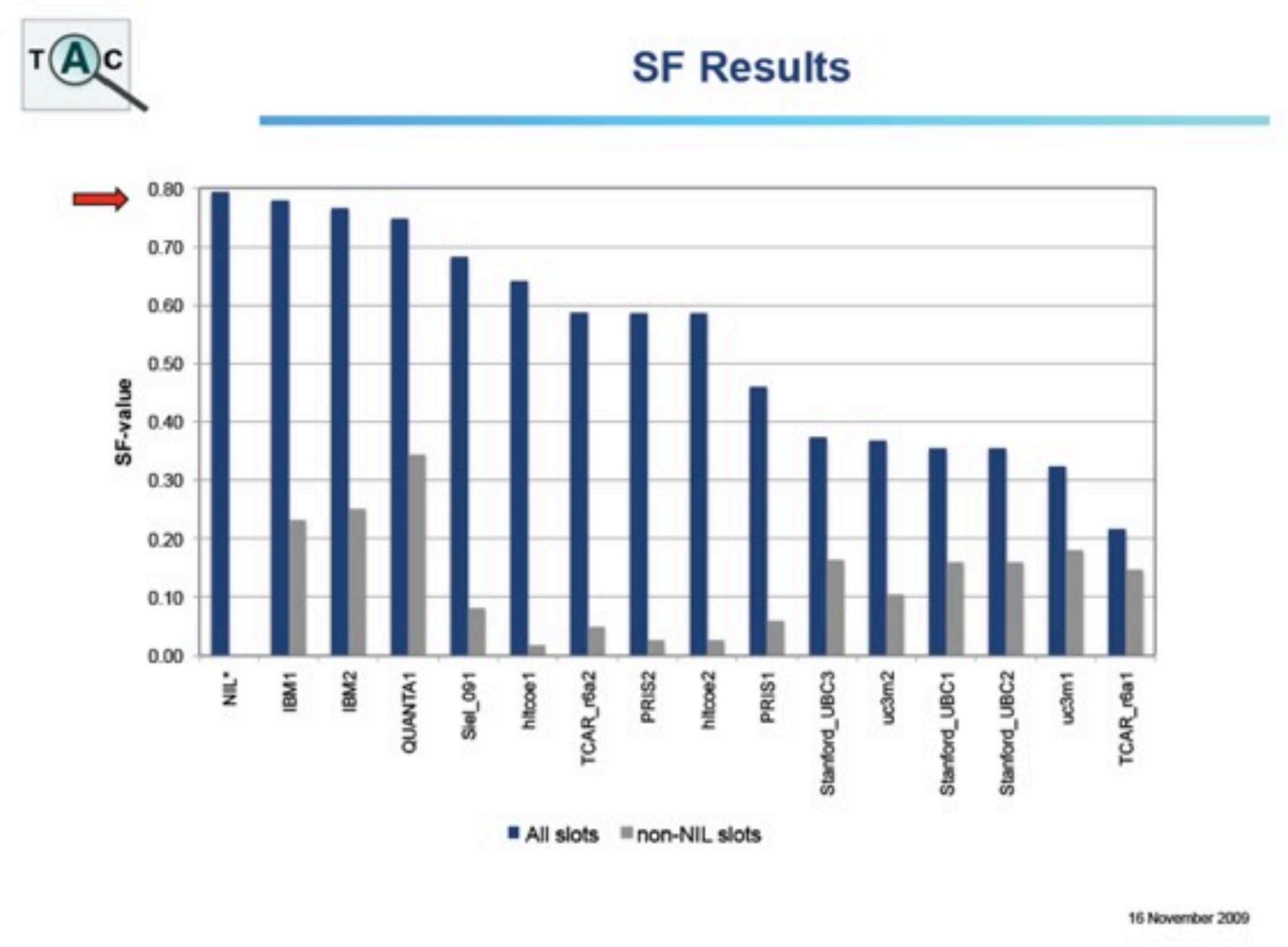
- **Date formed:** 12/2/1970
- **Website:** http://www.epa.gov/
- **Headquarters:** Washington, DC
- **Nicknames:** EPA, USEPA
- **Type:** federal agency
- **Address:** 1200 Pennsylvania Avenue NW

Optional: Also want to link some learned values within the KB:

- **Headquarters:** Washington, DC (kbid: 735)

16 November 2009

Textual Analysis Conference (since 2009) - Knowledge Base Population Track



Two Approaches to Building Extraction Systems

- Knowledge engineering approach
  - Grammars are constructed by hand
  - Domain patterns are discovered by a human expert through introspection and inspection of a corpus
  - Much laborious tuning and „hill climbing“
- Automatically Trainable Systems
  - Use statistical methods when possible
  - Learn rules from annotated corpora
  - Learn rules from interaction with user

Knowledge Engineering

- Advantages
  - With skill and experience, good performing systems are conceptually not hard to develop
  - The best performing systems have been hand crafted (still true for scenario patterns)
- Disadvantages
  - Very laborious development process
  - Domain adaptation might require re-configuration
  - Needs experts which have both, linguistic & domain expertise

Hand-Coded Methods

- Easy to construct in many cases
  - e.g., to recognize prices, phone numbers, conference names, etc.
- Easier to debug & maintain
  - especially if written in a “high-level” language (as is usually the case)
  - e.g.,

ContactPattern ← RegularExpression(Email.body,”can be reached at”)

PersonPhone ← Precedes(Person
Precedes(ContactPattern, Phone, D),
D)

- Easier to incorporate / reuse domain knowledge
- Can be quite labor intensive to write

Learning-Based Methods

- Can work well when training data is easy to construct and is plentiful
- Can capture complex patterns that are hard to encode with hand-crafted rules
  - e.g., determine whether a review is positive or negative
  - extract long complex gene names

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.“

- Can be labor intensive to construct training data
  - not sure how much training data is sufficient

Trainable Systems

- Advantages
    - Domain portability is relatively straightforward
    - System expertise is not required for customization
    - Data driven rule acquisition ensures full coverage of examples
- Disadvantages
    - Training data may not exist, and maybe very expensive to acquire
    - Large volume of training data may be required
    - Changes to specifications may require re-annotation of large quantities of training data

Possible solutions here are
- on-demand IE
- dynamic interactive IE
- ad-hoc IE

What works best?

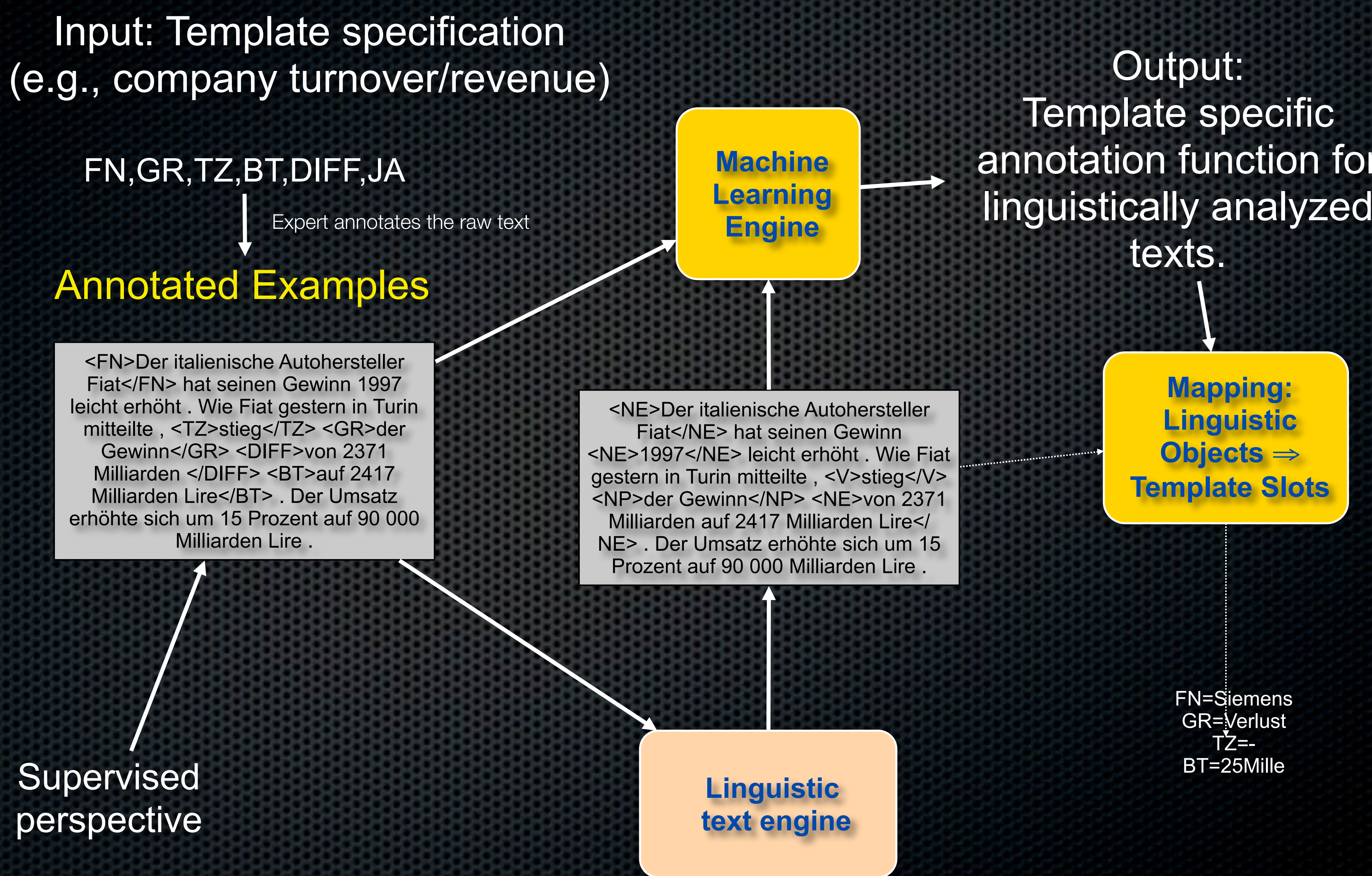
- Use rule-based approach when
  - Resources (e.g., lexicons, lists) are available
  - Rule writers are available
  - Training data scarce or expensive to obtain
  - Extraction specs likely to change
  - Highest possible performance is critical
- Use trainable approach when
  - Resources unavailable
  - No skilled rule writers are available
  - Training data is cheap and plentiful
  - Good performance is adequate for the task

This is still the main approach in industrial applications, where precision counts.

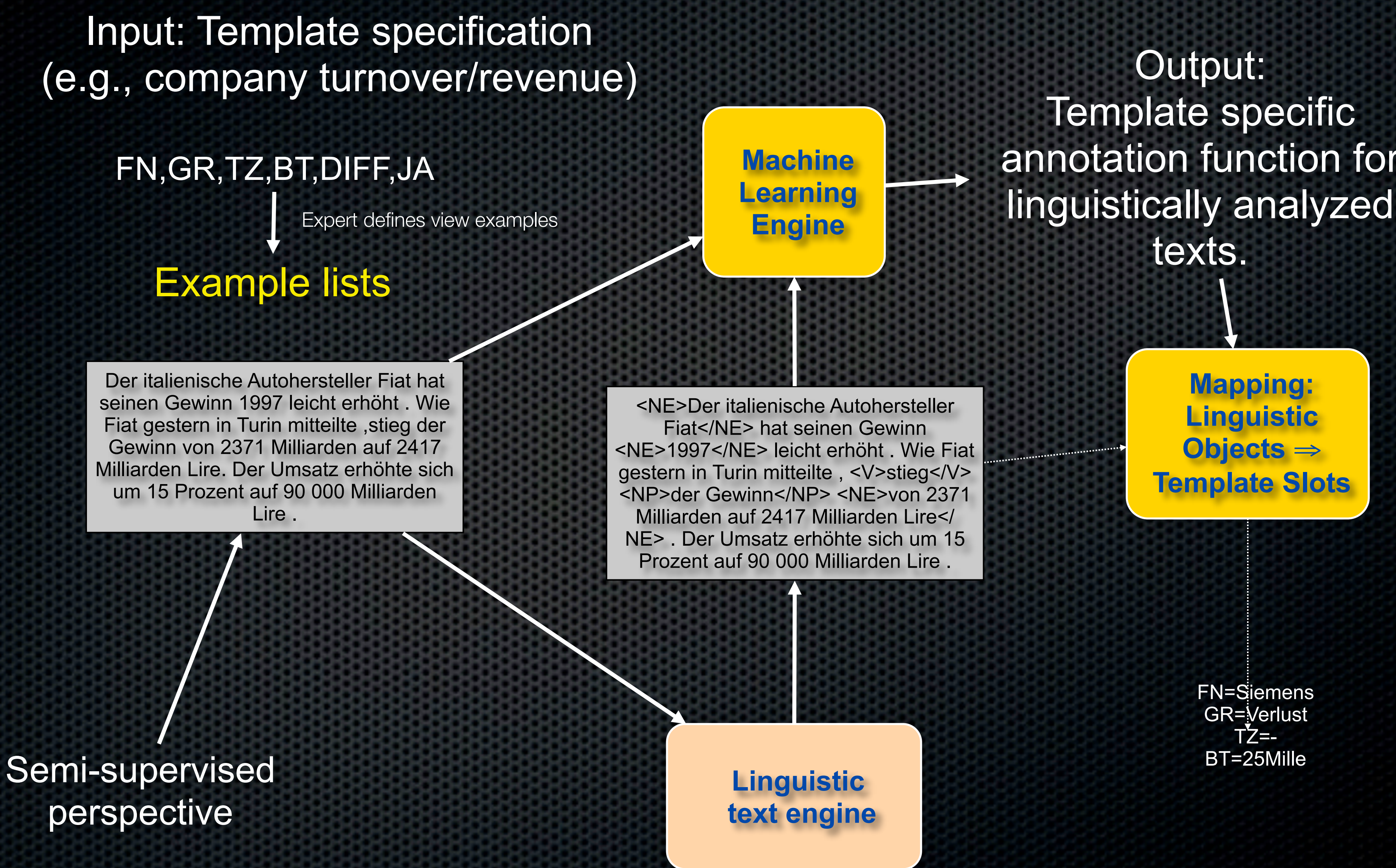
Architecture of Extraction Systems

- Domain-independent NL tools necessary
  - Main task: define scope of linguistic features
  - Major issue: robustness & efficiency
- Clean interface between domain-independent tools and domain-dependent
  - Domain modeling
  - Main Task: disambiguation
  - Easy adaptation of NL tools

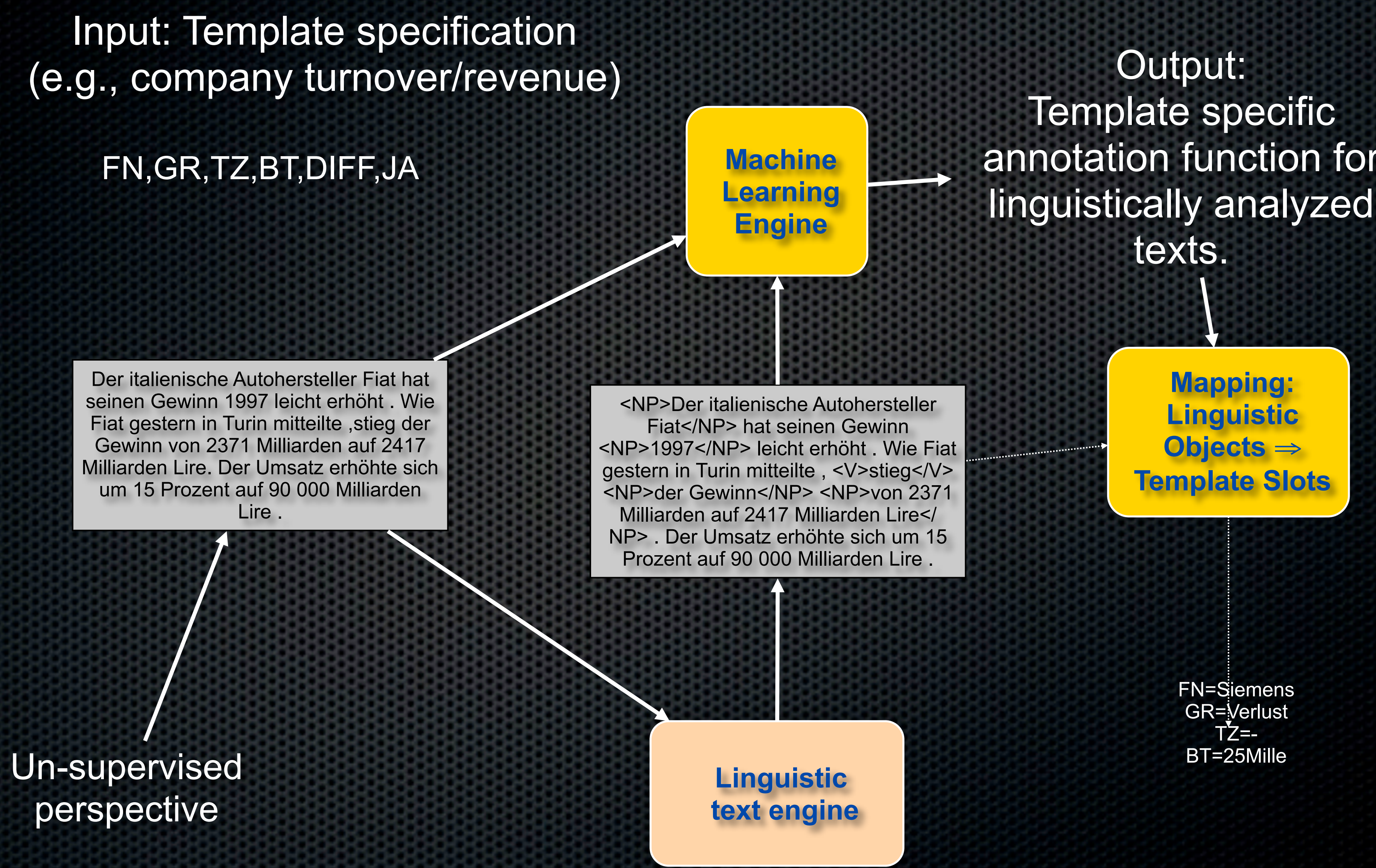
IE and Machine Learning



IE and Machine Learning



IE and Machine Learning



In any case, major subtasks

- Linguistic feature extraction
- Recognition and extraction of Named Entities
- Recognition and extraction of Relations
- Recognition and extraction of Events

About the Linguistic Core Engine

- From part-of-speech, to morphology, to phrase recognition, to full sentence structure to logical formulas.
- What linguistic features are needed or useful depends on the application task and strategy.

General Framework of NLP

John runs.

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing
Interpretation

General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

General Framework of NLP

John runs.

John run+s.

P-N V 3-pre
N plu

Morphological and Lexical Processing

⇩

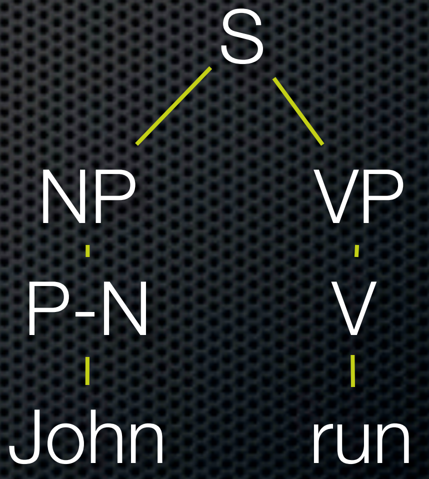
Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation



General Framework of NLP

John runs.

Morphological and Lexical Processing

⇩

John run+s.
P-N V 3-pre
N plu

Syntactic Analysis

⇩

S
NP VP
P-N V
John run

Semantic Analysis

⇩

Pred: RUN
Agent:John

Context processing
Interpretation

General Framework of NLP

John runs.

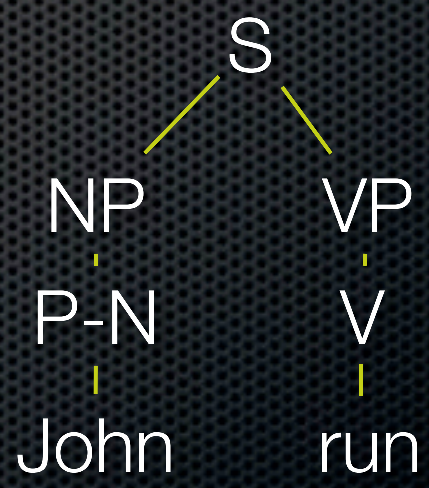
Morphological and Lexical Processing

John run+s.

P-N V 3-pre

N plu

Syntactic Analysis



Semantic Analysis

Pred: RUN

Agent:John

Context processing Interpretation

John is a student.

He runs.

Difficulties of NLP

(1) Robustness: Incomplete Knowledge

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Incomplete Lexicons
Open class words
Terms
Term recognition
Named Entities
Company names
Locations
Numerical expressions

Difficulties of NLP

(1) Robustness: Incomplete Knowledge

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing Interpretation

Incomplete Grammar

Syntactic Coverage

Domain Specific

Ungrammatical

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP

Morphological and
Lexical Processing

Syntactic Analysis

Semantic Analysis

Context processing
Interpretation

Predefined
Aspects of
Information

Incomplete
Domain Knowledge
Interpretation Rules

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

(2) Ambiguities:
Combinatorial
Explosion

General Framework of NLP

Morphological and Lexical Processing

⇩

Syntactic Analysis

⇩

Semantic Analysis

⇩

Context processing
Interpretation

Most words in English
are ambiguous in terms
of their part of speeches
runs: v/3pre, n/plu
clubs: v/3pre, n/plu
and two meanings

Difficulties of NLP

(1) Robustness:
Incomplete Knowledge

General Framework of NLP

(2) Ambiguities:
Combinatorial
Explosion

Morphological and Lexical Processing

↓

Syntactic Analysis — Structural Ambiguities

↓

Semantic Analysis — Predicate-argument Ambiguities

↓

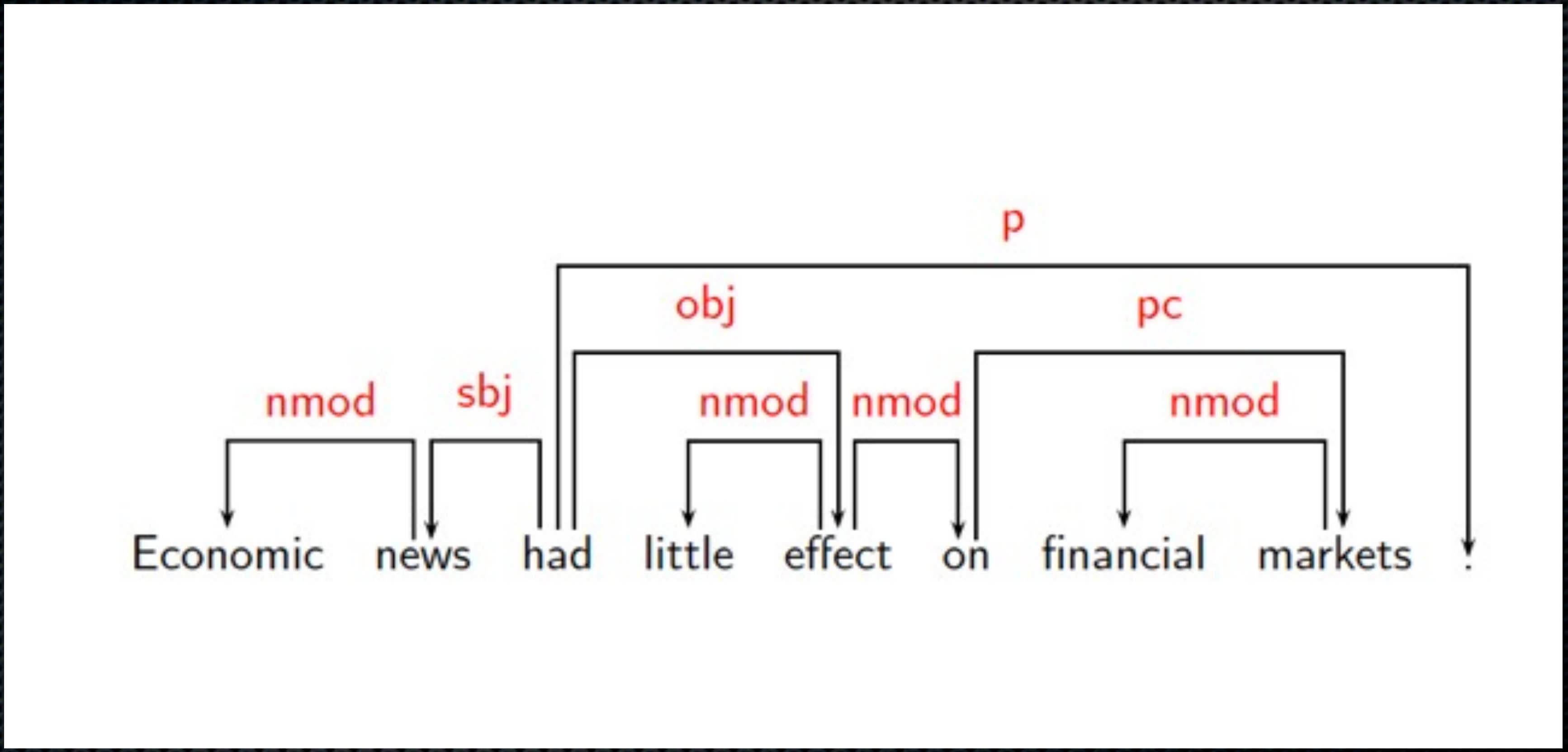
Context processing Interpretation

Combinatorial Explosion

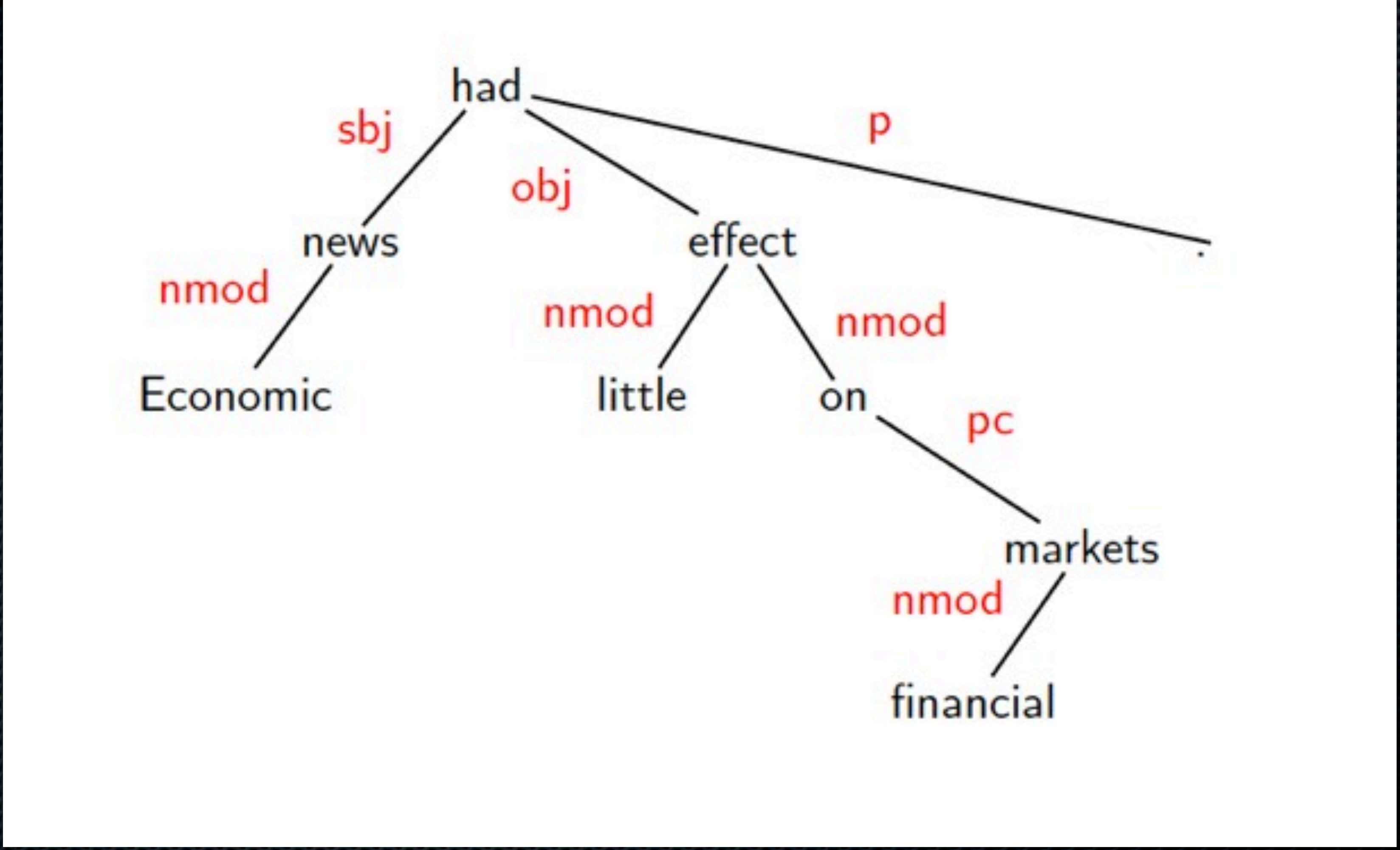
Dependency Parsing

- The basic idea:
  - Syntactic structure consists of lexical items , linked by binary asymmetric relations called dependencies
  - The dependency structure of a sentence is a acylic directed graph (DAG; mostly, they are such trees)
- A dependency parsing
  - computes a dependency structure for a sentence
  - tree-bank-based dependency parsers: the rules or constraints for determining the valid structure is learned from a large set of example valid dependency structures (trees)

Dependency Structure



Notational Variants

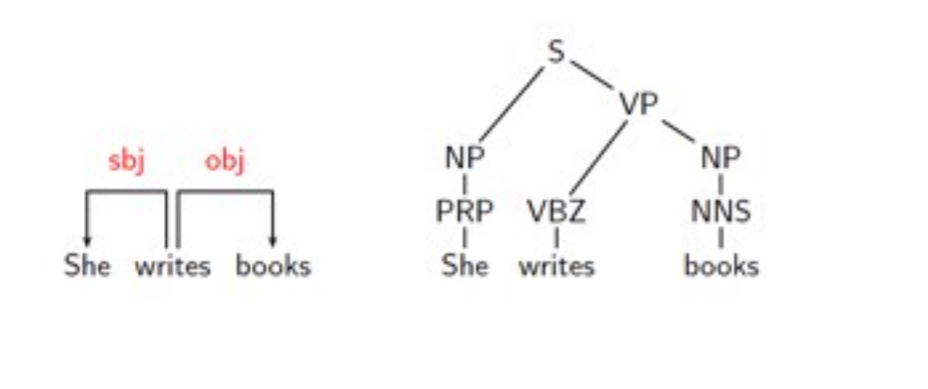


Comparison

- Dependency structures explicitly represent
  - head-dependent relations (directed arcs),
  - functional categories (arc labels),
  - possibly some structural categories (parts-of-speech).
- Phrase structures explicitly represent
  - phrases (nonterminal nodes),
  - structural categories (nonterminal labels),
  - possibly some functional categories (grammatical functions).

Application-oriented Advantage

- Shallow semantics: Direct encoding of predicate-argument structure
- Machine Learning:
  - feature extraction
  - Tree kernels



IE: compromise NLP (cf. Appelt & Israel, 1999; Neumann & Piskorski, 2002)

- Task characteristic
  - Lots of texts
  - Dirty texts
  - World knowledge needed

- Compromise
  - Finite-state models
  - Robust techniques
  - Domain specific processing at each stage of analysis

The bottom line:
Find the most favorable tradeoff between recall and precision for the task at hand.

Languages Are Structural

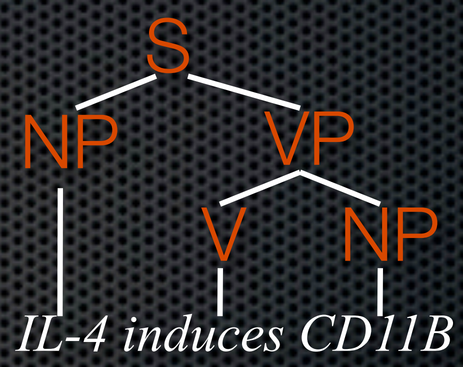
governments
Firmen

Slide from Hoifung Poon, Dept. of Computer Science & Eng., University of Washington, NAACL tutorial, 2010

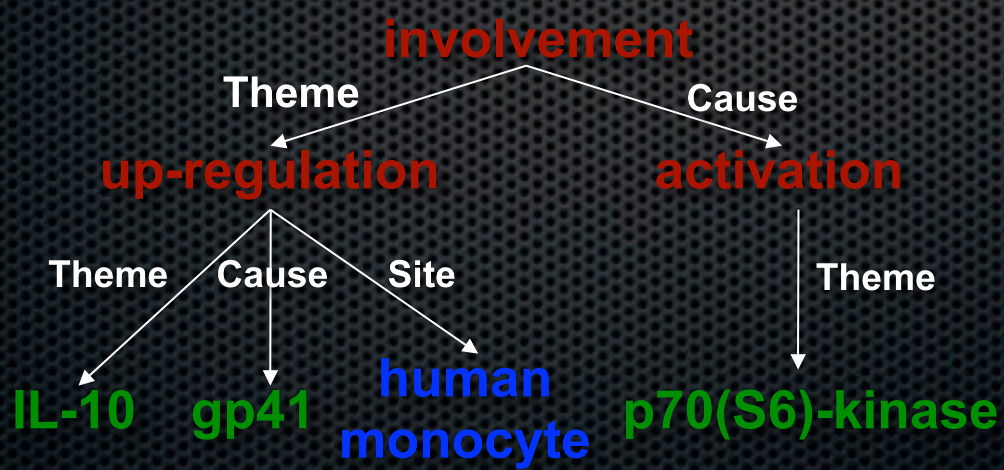
Languages Are Structural

govern-ment-s

Firma-PL



Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41……



George Walker Bush was the 43rd President of the United States.

……

Bush was the eldest son of President G. H. W. Bush and Babara Bush.

…….

In November 1977, he met Laura Welch at a barbecue.

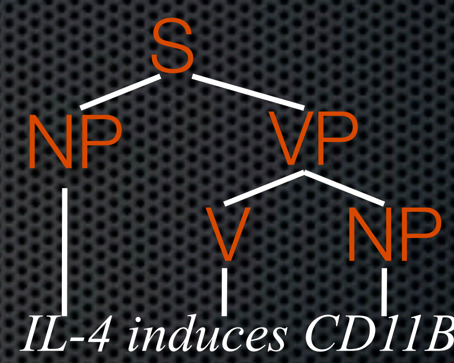
Slide from Hoifung Poon, Dept. of Computer Science & Eng., University of Washington, NAACL tutorial, 2010

Languages Are Structural

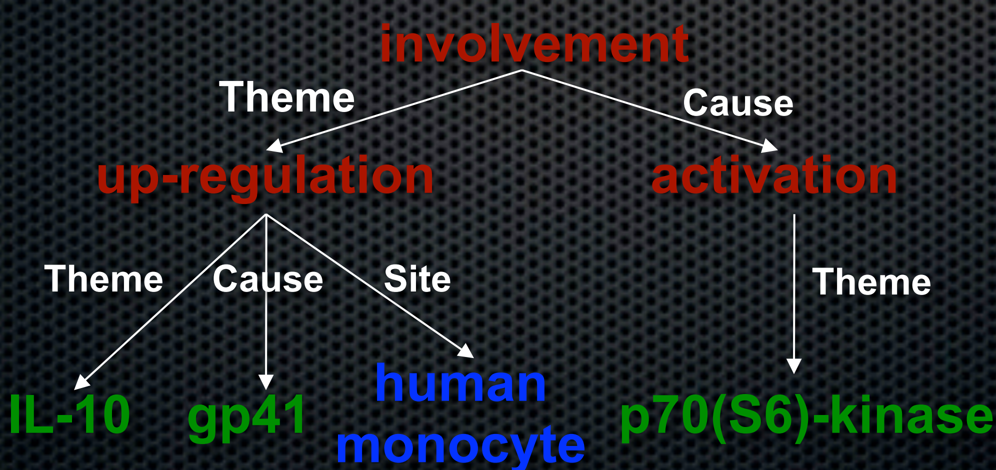
govern-ment-s

l-m$px-t-m

(according to their families)



Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41……



George Walker Bush was the 43rd President of the United States.

……

Bush was the eldest son of President G. H. W. Bush and Babara Bush.

…….

In November 1977, he met Laura Welch at a barbecue.

Slide from Hoifung Poon, Dept. of Computer Science & Eng., University of Washington, NAACL tutorial, 2010

Languages Are Structural

- Objects are not just feature vectors
  - They have parts and subparts
  - Which have relations with each other
  - They can be trees, graphs, etc.
- Objects are seldom i.i.d.
  (independent and identically distributed)
  - They exhibit local and global dependencies
  - They form class hierarchies (with multiple inheritance)
  - Objects' properties depend on those of related objects
- Deeply interwoven with knowledge

Slide from Hoifung Poon, Dept. of Computer Science & Eng., University of Washington, NAACL tutorial, 2010

Languages Are Statistical

I saw the man with the telescope

NP

I saw the man with the telescope

NP ADVP

I saw the man with the telescope

Microsoft buys Powerset

Microsoft acquires Powerset

Powerset is acquired by Microsoft Corporation

The Redmond software giant buys Powerset

Microsoft's purchase of Powerset, …

Here in **London**, Frances Deek is a retired teacher …
In the Israeli town …, Karen **London** says …
Now **London** says …

London = PERSON or LOCATION?

G. W. Bush ……
…… **Laura Bush** ……
Mrs. Bush ……

Which one?

Languages Are Statistical

- Languages are ambiguous
- Our information is always incomplete
- We need to model correlations
- Our predictions are uncertain
- Statistics provides the tools to handle this

Slide from Hoifung Poon, Dept. of Computer Science & Eng., University of Washington, NAACL tutorial, 2010

NL and Information Extraction

- Text ambiguity and variability is a major challenge for large-scale robust and efficient text exploration
- One has to carefully decide, which linguistic properties are actually relevant and necessary for high-coverage and high-precise extraction of information nuggets, e.g., named entities or relations between them.